White Paper Report

Report ID: 104050

Application Number: HG5001009

Project Director: Helen Aristar-Dry (hdry@linguistlist.org)

Institution: Eastern Michigan University

Reporting Period: 7/1/2009-6/30/2012

Report Due: 9/30/2012

Date Submitted: 9/27/2012

White Paper

Grant Number: HG5001009

RELISH: Rendering Endangered Language Lexicons Interoperable Through Standards Harmonization

DFG/NEH Bilateral Digital Humanities Program

DFG: Primary Project Director: Prof. Jost Gippert, Universität Frankfurt

NEH: Primary Project Director: Prof. Helen Aristar-Dry, Eastern Michigan University

Submitted September 27th, 2012

Throughout the course of the RELISH project, the Max Planck Institute for Psycholinguistics (MPI), the Johann Wolfgang Goethe-Universität Frankfurt, and LINGUIST List at Eastern Michigan University (EMU) developed tools to aid in the harmonization of digital standards for lexical information in Europe and America, and in the creation of an interchange format to interoperate among existing lexicons of endangered languages. During the three years of RELISH project work, the team successfully harmonized terminology between the General Ontology for Linguistic Description (GOLD) and the International Standards Organization Data Category Registry (ISOCat). The team also harmonized three distinct data structures: GOLD, Lexical Interchange Format (LIFT), and the Lexical Markup Framework (LMF) core features plus extensions. To do this, the team processed lexica of the under-documented languages (Udi, Wichita, and Tuva) and include these languages in the Lexicon Enhancement via the GOLD Ontology (LEGO) database.

Complete harmonization of LIFT with the LMF core features was delayed by extensive revision of the RELISH interchange format schema. These revisions were made through extensive correspondence between LINGUIST List and MPI. Additionally, the composition of stylesheets used in the interchange process (also known as 'round-tripping') required integration of the LEGO project export format, which was also undergoing extensive development during the RELISH project.

The goal to extend existing standards for managing semantic domains and mapping them to external anchors was also delayed. The participants at LINGUIST List and at the MPI discussed multiple proposals to extend these standards, but did not have the opportunity to implement any one proposal. The team did, however, decide to either make use of concepticons already in use by the LEGO project, or create self-documenting data categories for every element. The schema could map these semantic domains to WordNet, a network of meaningfully related words and concepts, thus extending the scope of the RELISH project outcomes

Several difficulties were faced in conducting work on the project. Communication between all participants was difficult at times due to the differences in time zones and participant availability. However, correspondence was managed through extensive emails, as well as Skype conferences and phone calls when major updates needed to be discussed.

There also existed a disparity of encoding standards between LIFT and LMF XML. Because of this, the team had to evaluate features on a case by case basis, and eventually decided to incorporate elements from LIFT, LMF, and the Text Encoding Initiative (TEI) in order to both maintain LMF compliance and to include all relevant lexical information.

Issues with GOLD proved problematic to the RELISH project. First, the Ontology required extensive cleanup, which included adding, removing, and editing definitions of linguistic terminology. Also, relevant sources needed to be added to these definitions. These changes to GOLD resulted in a more useful ontology for the purposes of the RELISH project, and for more general scholarly use.

Despite various staff changes at MPI and LINGUIST List, handovers were smooth and resulted in little or no delay of progress on project due to thorough documentation.